UNIVERSITY OF CALIFORNIA SAN DIEGO

Community-level Monitoring of HIV Spread

A thesis
submitted in partial satisfaction of the
requirements for the degree
Master of Science
in
Electrical and Computer Engineering

by

Sina Malekian

Committee in charge:

Professor Siavash Mir Arabbaygi, Chair
Professor Paul Siegel
Professor Behrouz Touri

2020

The thesis of Sina Malekian is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California San Diego

2020

Dedication

This is dedicated to my family for their unconditional love and support.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor, Professor Siavash Mir Arabbaygi for his knowledge and guidance throughout my program. I would also thank my family and friends who made this possible.

ABSTRACT OF THE THESIS

Community-level Monitoring of HIV Spread Prevention

by

Sina Malekian
Master of Science in Electrical and Computer Engineering
University of California San Diego, 2020
Professor Siavash Mirarab Baygi, Chair

Health departments are using HIV data to monitor HIV growth in real time. The main purpose of this monitoring is to come up with policies for efficient allocation of medical resources. In order to achieve the efficient medical resources allocation, a method should be established for predicting where future transmissions of HIV will occur using the partial information of the transmission history. Validity of these predictions are of paramount importance as it affects the policy for allocation of medical resources. Indeed, the more accurate the prediction is, the more efficiently preventive care or other resources can be allocated to the network. The focus of this work is on community-level monitoring of HIV spread prevention. We have modeled the sexual network as communities of individuals and proposed community-level methods for prediction. Then, we have compared predictive power of the proposed methods in different settings of the network.

Chapter 1

Introduction

The HIV spread process in social and sexual network has been widely studied to pose interventions and reduce risk for HIV infecion [1, 2]. Previous studies support the idea of transmission history reconstruction for HIV spread prevention. For instance, the statistical power of network-based statistics have been studied to measure and investigate the treatment options to reduce the future transmissions of HIV [3]. Sexual contact network reconstruction will provide valuable insight about the spread of HIV [4]. Computational approaches have been developed to estimate the HIV-1 evolutionary rates with transmission data reconstruction [5]. Phylogenetic inference methods have been deployed to reconstruct the transmission history [6].

The captivating question here is whether clusters of HIV transmissions can be monitored in real time. The goal would be to reconstruct the transmission network among HIV+ individuals, real-time monitor the growth of reconstructed clusters, and identify the

top growing clusters. The real time monitoring of HIV clusters will enable use to propose methods to allocate the medical resources to the growing clusters. This process is referred as cluster size monitoring (CSM). Methods have been implemented to identify the individuals who are highly related to the transmission events and pose targeted interventions [7]. Structures of local epidemic can be reconstructed and methods have been implemented to design and evaluate control interventions [8]. HIV-1 transmission network has been reconstructed to evaluate efforts to prevent future transmissions [9].

The common assumption here is that the past transmission history gives us information about the future transmission. To be more elaborate, identifying the top growing clusters of diagnosed individuals will give us information about the clusters which are likely to have new infections. This assumption requires evaluation and we are not the first to notice this assumption. Previous studies have pinpointed to this assumptions and suggested that the common assumptions need to be evaluated [10][11][12].

The innovation of our work is to evaluate the usefulness of CSM in different settings of the network. The goal is to deploy a real-time CSM system with varying parameters of sexual network and evaluate the predictive power of this method in different settings of the network. In order to formulate and evaluate the predictive power of CSM, it has been deployed on a sexual network consisting of communities of individuals with fixed sizes. The reason behind this is that the notion of the cluster is not well-defined and different methods have been implemented to infer transmission clusters [13][14][15][16]. We desire to have a robust framework for evaluation of the predictive power of growth-based approach (not the accuracy of clustering method) to identify the top growing communities which are likely to have future transmissions [17]. We refer to this real-time monitoring method as

community-level monitoring of HIV spread.

We will provide a framework to judge the predictive power of the growth-based approach [17] for community-level monitoring. The key question here is whether monitoring the diagnosed cases in the communities will be useful to identify the communities which are likely to have new infections. The goal is to predict future infections and not new diagnoses. In order to measure the predictive power of community-level monitoring, we simulate the transmission scenario from the time $(t - \Delta t, t + \Delta t)$. Then, based on the transmission events in the interval $(t - \Delta t, t)$, communities will be sorted with respect to their growth rate of diagnosed cases proposed by growth-based approach [17]. Then, the communities will be sorted with respect to their future growth rate of infected cases in the time interval $(t, t + \Delta t)$. The correlation between the two orderings can be measured as predictive power of the proposed community-level monitoring method.

Regarding the difficulty of evaluation in clinical studies, we have used a simulation framework called FAVITES [18]. FAVITES is a modular framework for simulating the epidemic in different settings of the network. The simulation procedure consists of series of interactions between abstract modules. Implementation of modules is user-specified in the sense that users can specify the implementation regarding the context for which they are applying FAVITES. In the simulations, SIR-like HIV-specific models of transmission have been implemented [19][20][21].

Given the potential time lag between diagnosis and infection of the individuals, time to diagnosis is an important parameter in monitoring the HIV-growth in the community level. Our work thoroughly explores the impact of this potential time lag on the predictive

power of the growth-based approach. Also, we have proposed a mathematical method for estimation of number of infected cases based on the number of diagnosed cases and network parameters. We will also evaluate the effectiveness the proposed approach for modifying the time lag between diagnosis and infection.

Chapter 2

Prioritization Problem

2.1 Problem Statement and Formulation

In our model, we have a contact network consisting of multiple communities in which HIV transmissions will occur. A community is defined as a sexual network of individuals with a high probability of sexual contact. More precisely, the contact network is assumed to consist of communities with high probability of sexual contact within the community. However, an inter-community sexual contact might occur with a lower probability. The transmission procedure is initiated at time zero with certain number of seeds (individuals who are initially infected) and spreads throughout the network. The main question here is how accurately the future HIV transmissions can be predicted, and how reliable the predictions are in different settings of the network.

Prioritization problem in this context can be defined as finding the top growing com-

munities among all of the communities in the network. The input of this problem is the imperfect information from HIV transmission history. The output is an ordering of communities based on their growth rate.

Suppose that we have M communities in the network each of which has N individuals. We model the the HIV spread as follows. The process starts at time zero and continues until all of the people in our network are infected. Each of the communities have a number of infected and diagnosed people at any point of time during the HIV spread. The ultimate goal is to come up with a method to sort these M communities based on their growth rate.

We will denote the number of infected and diagnosed individuals in community c at time t with $D_c(t)$ and $I_c(t)$ respectively. It is assumed that we know $D_c(t)$ at any point of time for all of the communities in our network. However, $I_c(t)$ is not known. The implications of this lack of information will be discussed in the following chapters.

The formulation of prioritization problem is stated as below:

Suppose that we have communities $C_1, C_2, ..., C_M$ in our network. Given $D_c(t)$ as the number of diagnosed people at time t for community c, sort the communities based on predicted future growth. The output ordering will be $S = (S_1, S_2, ..., S_M)$, which is a permutation of $C_1, C_2, ..., C_M$ communities. The identified top-growing communities should ideally have the most number of future transmissions per infected individual.

Therefore, all of the methods to solve this prioritization problem should come up with a metric $G(t)$ to sort the communities which predicts the future growth of the communities.

2.2 Challenges

As stated in the previous section, the prioritization problem is defined as the problem of prioritizing all communities of our network at any point of time during HIV spread procedure. There are several challenges for designing methods to solve this problem. The challenges will be elaborated on in what follows.

Prediction. The perfect real-time monitoring of HIV spread in the network requires the knowledge from future. However, the information that we have from our transmission network corresponds to the past events not the future events. For instance, we know the number of diagnosed people at any point of time in the past but there is no certain information about the number of diagnosed people in the future. Consequently, we should be able to extract useful information from past transmission history, network settings, and dynamic of the HIV spread to address this challenge. Our proposed method is designed to address this challenge based on a metric for growth rate which is defined as the rise in the number of diagnosed people in the time slot $(t - \Delta t, t)$ such that t indicates the current time in HIV spread process.

Missing information. Furthermore, the full transmission history of the network is not accessible. One of the main issues in the second challenge is that we only know the time that each individual is diagnosed not the time that the individual is infected. Consequently, $D_c(t)$ is assumed to be known for all of the communities in our network. However, there will be no information on $I_c(t)$ which is the number of infected people at time t in community c. Therefore, the proposed method should be able to deal with the inaccuracy caused by the

lack of information. In this work, we have come up with a correction algorithm to address this issue. The method is designed to estimate the number of infected people denoted as $I_c(t)$ using the number of diagnosed people $D_c(t)$ and certain parameters of the network (assumed to be known).

2.3 Growth-based Ordering Approach

As discussed earlier, any approach for solving prioritization problem will come up with metric $G(t)$ for sorting the communities. The metric will be referred to as the growth rate of each community. In the growth-based ordering approach, $G(t)$ is defined as the relative number of recent past transmissions to the overall number of diagnosed individuals in each community [17].

Let $D_c(t)$ and $I_c(t)$ denote the number of diagnosed and infected people from community c at time t, respectively. The metric $G(t)$ in the growth-based approach is defined as:

$$G_c^D(t-\Delta t, t) = \frac{D_c(t) - D_c(t-\Delta t)}{D_c(t-\Delta t)} \tag{2.1}$$

where the subscript c refers to the community c which is the community for which the growth rate G_c^D is being computed. The superscript D in G_c^D indicates that the growth-based approach will make use of information about the number of diagnosed people in each community, and there will be no information on the number of infected people. The growth rate can also be defined for infected people in community c as below:

$$G_c^I(t-\Delta t, t) = \frac{I_c(t) - I_c(t-\Delta t)}{I_c(t-\Delta t)} \tag{2.2}$$

where subscript c refers to the community c and subscript I indicates the fact that the growth rates are computed using the number of infected individuals. However, as noted earlier, the growth rate for infected people cannot be used as the infection time for individuals is not known in reality.

Both metrics show the average number of new infected (diagnosed) individuals normalized by the previously infected (diagnosed) individuals between time t and $t - \Delta t$. We will then sort the communities with respect to their diagnosis growth rate. Input of prioritization problem is the transmission history and transmission events in the network without knowing the infection time of the individuals. The output ordering S of the problem is an ordering of communities based on their growth rate at any given time.

Both of these growth-based metrics are based on a fundamental assumption: that the rate of per-person transmission in the near future will be similar to the rate of per-person transmission in the near past. This assumption will be not always be correct, especially if the time period defining "near" is allowed to be long. The per-person rate of transmission can slow down or increase, and these changes are ignored when we order the communities based on the past growth rate.

We can say intuitively that if we could access the infection time for individuals, our estimation for the growth rate of each community would be more accurate. As a result, we would expect to identify the top-growing communities more accurately when we use infection time comparing to the accuracy resulted from using diagnosis time.

2.4 Summary

Prioritization problem is defined as finding the top growing communities among all of them in HIV spread process. Input of the prioritization problem is the imperfect transmission history. The goal is to give an ordering of communities with respect to their future rate of transmission. Computing perfect ordering is challenging due to several reasons, including the need for predict future transmissions and an imperfect knowledge of the past. We will make an attempt to address these challenges in our growth-based ordering approach presented in the next chapters.

In the growth-based approach to prioritization, a metric is introduced to sort the respective communities. This metric relies on the number of diagnosed (alternatively, infected) people in the recent past to estimate the current growth rate of each of the communities. Then, based on the computed growth rate for each community, there will be an ordering of communities which is the output of prioritization problem.

Chapter 3

Correction for Diagnosis Lag

3.1 Problem Statement

As it is stated in the prioritization problem, we only know the diagnosis time for HIV transmissions. One of the main challenges for solving prioritization problem is the inaccurate information from past transmission history. There is always a time lag between HIV diagnosis and HIV infection, and this lag between diagnosis and infection may cause significant inaccuracy for solving prioritization problem. Indeed, we can only use the number of diagnosed people in our calculation to identify top-growing communities not the number of infected people. Consequently, all of the accessible information of the network is subject to a time lag between diagnosis and infection.

We conducted a simulation to illustrate the HIV spread through a single community of $N = 500$ individuals. Simulation starts at time zero and continues until all of the

individuals are infected. Figure 3.1 represents the growth in the number of diagnosed and infected individuals in the community. Previous work has been conducted to model the curves of infected and diagnosed individuals over time as the sigmoid functions [22]. As it can be observed in the figure, both of diagnosis and infection curves are very much similar to the sigmoid functions. The number of infected and diagnosed people increase slowly at the beginning of epidemic. However, as the number of infected people increases, we observer a rise in the number of transmissions and the number of infected people in the community. Eventually, the number of infected people is saturated since there will not be many uninfected individuals in the community unlike the beginning of epidemic.

In the following sections, we have exploited this property and modeled the infection and diagnosis curve as sigmoid functions to modify the lack of information caused by the time lag between infection and diagnosis.

In order to illustrate the lack of information, we delve into an example. Consider two communities with corresponding infection and diagnosis curves in Figure 3.2. Recall the prioritization problem for these two communities between $t = 9$ and $t - \Delta t = 8.5$. Assume that we have the full information, including $I_C(t)$. If the growth-based ordering approach of the infection curve is taken into account for the prioritization problem:

$$G_1^I(t - \Delta t, t) = \frac{I_1(9) - I_1(8.5)}{I_1(8.5)} = 0.074$$

$$G_2^I(t - \Delta t, t) = \frac{I_2(9) - I_2(8.5)}{I_2(8.5)} = 0.198$$

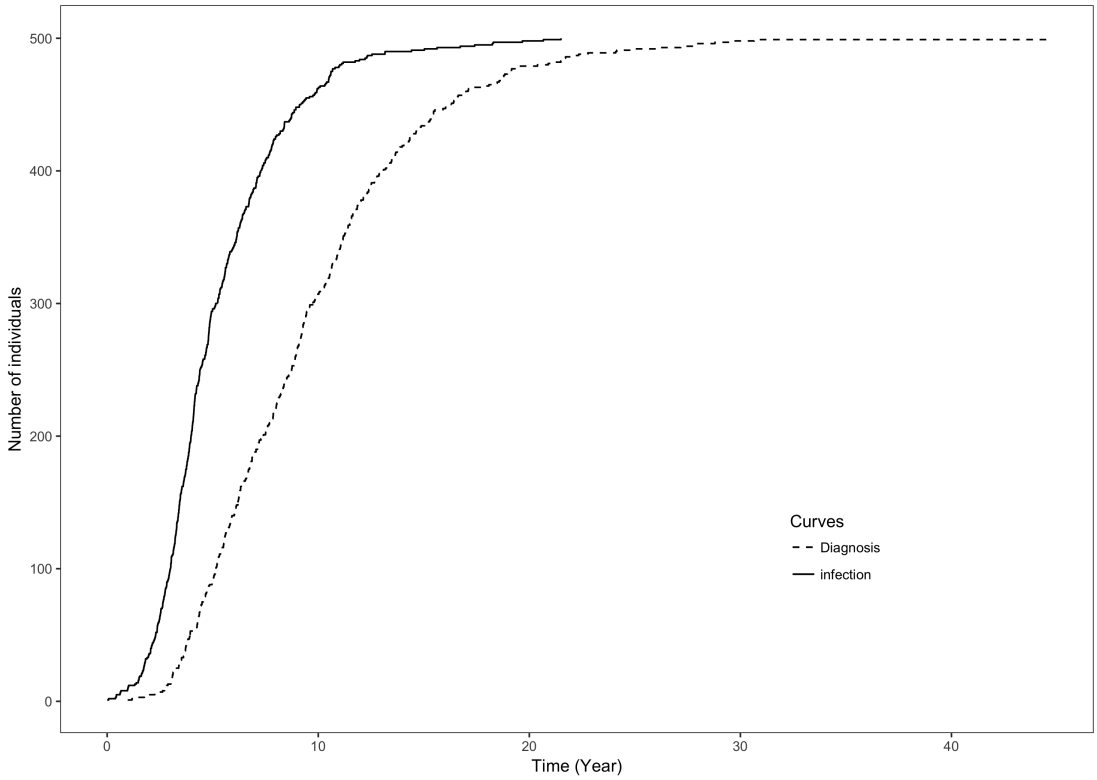

Figure 3.1. Number of diagnosed and infected individuals in a single community with $N = 500$ individuals. Simulation starts at time zero and continues until all of the individuals in the community are infected.

Conversely, if the diagnosis growth-based approach is applied to this prioritization problem, the diagnosis growth rate of the two communities will be computed as below.

$$G_1^D(t-\Delta t,t)=\frac{D_1(9)-D_1(8.5)}{D_1(8.5)}=0.49$$

$$G_2^D(t-\Delta t,t)=\frac{D_2(9)-D_2(8.5)}{D_2(8.5)}=0.4$$

Then, if the growth-based approach is deployed taking into account of the number of infected individuals, community 2 should be prioritized to community 1 as $G_1^I(8.5,9) < G_2^I(8.5,9)$. However, the number of infected individuals is not known in reality. Therefore, if instead of infection time we exploit diagnosis time, community 1 would be prioritized to community 2 as $G_1^D(8.5,9) > G_2^D(8.5,9)$.

As it is noticeable in the previous example, the time lag between diagnosis and infection may cause inaccuracy for sorting the communities based on their growth rate. Since there the number of infected individuals in the community is not directly observed, the interesting question here is whether it would be possible to present a method to estimate the number of infected people based on the number of diagnosed people and network features. Then, the inaccuracy caused by the time lag may be addressed to some extent.

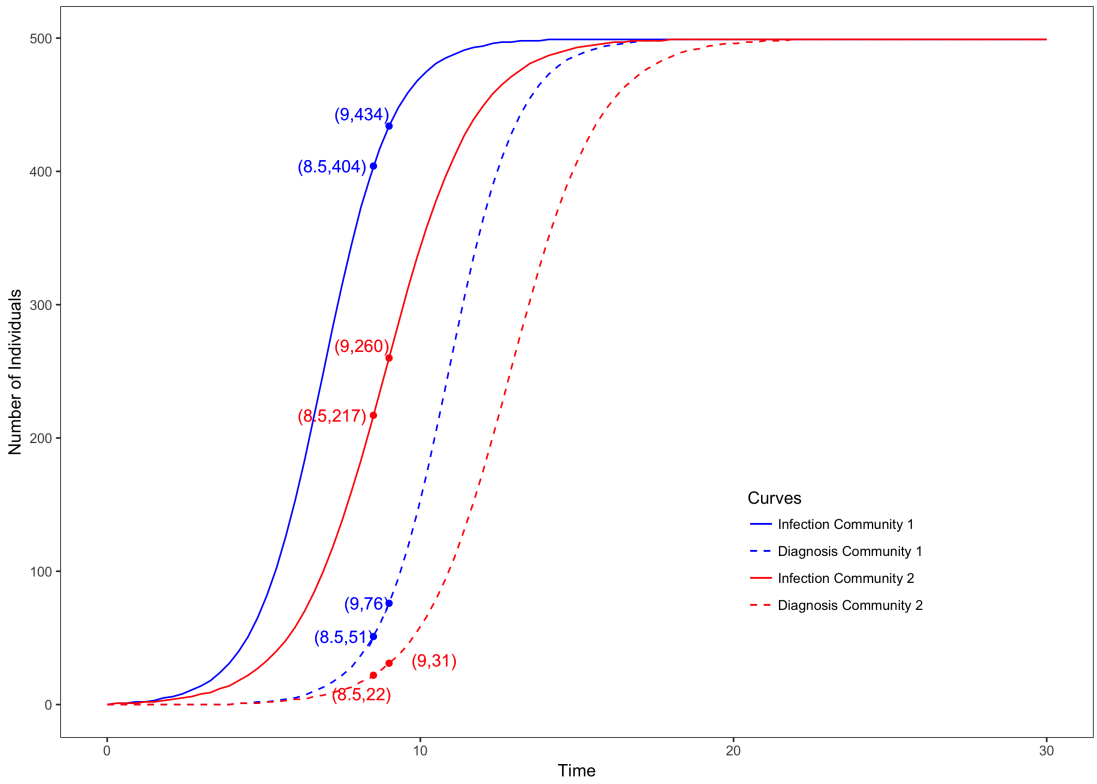

Figure 3.2. Number of diagnosed and infected individuals in two communities with $N = 500$ individuals. HIV spread starts at time zero and continues until all of the individuals in the communities are infected. This figure demonstrate the inaccuracy caused by the time lag between infection and diagnosis for solving the prioritization problem.

3.2 Correction Algorithm

The key concept of this section is to estimate the number of infected people in each community using the number of diagnosed people which would hopefully lead to a more accurate ordering for top-growing communities.

We also assume that we have access to two parameters:

- the population of the communities (i.e., the number of susceptible people in the community), and
- the expected time from infection to diagnosis across all individuals in our network.

Assume that the number of infected and diagnosed people in a specific community are $i(t)$ and $d(t)$ respectively. Let's denote the number of individuals in the community as N, the expected time from infection to diagnosis as T, and the duration of time span from the first diagnosed individual until the present time as t_d.

We approximate both functions of the number of diagnosed and infected people versus time with sigmoid functions. We assume the SIR model and under a more restrictive SIR model, sigmoid functions are proved to be the limiting behavior. Further, we make a second strong assumption: the two sigmoid functions relate to each other with a time lag equal to T. Given this simplifying assumption, the following equations can be obtained for i and d:

$$\hat{d}(t) = \frac{N}{1 + (N-1)e^{-rt_d}} \tag{3.1}$$

$$\hat{i}(t) = \frac{N}{1 + (N-1)e^{-r(t_d+T)}} \tag{3.2}$$

Thus, we can rewrite the equations as below:

$$t_d = \frac{-1}{r} \ln(\frac{1}{N-1}(\frac{N}{\hat{d}} - 1)) \tag{3.3}$$

$$t_d + T = \frac{-1}{r} \ln(\frac{1}{N-1}(\frac{N}{\hat{i}} - 1)) \tag{3.4}$$

If the equation 3.4 is divided by the equation 3.3, we will have the following equation:

$$1 + \frac{T}{t_d} = \frac{\ln(\frac{1}{N-1}(\frac{N}{\hat{d}} - 1))}{\ln(\frac{1}{N-1}(\frac{N}{\hat{i}} - 1))} \tag{3.5}$$

Thus,

$$\frac{1}{N-1}(\frac{N}{\hat{i}} - 1) = (\frac{1}{N-1}(\frac{N}{\hat{d}} - 1))^{(1+\frac{T}{t_d})} \tag{3.6}$$

As we know the true value for the number of diagnosed individuals, we can simply estimate $\hat{d}$ as the number of diagnosed people which is assumed to be known at any point of time. Thus, we can have a estimation of the number of infected individuals using the equation 3.8 which is derived from equation 3.6:

$$\hat{i} = \frac{N}{1 + (N-1)(\frac{N-\hat{d}}{\hat{d}(N-1)})^{(1+\frac{T}{t_d})}} \tag{3.7}$$

To wrap up the algorithm, the correction algorithm is designed to address the imperfect information challenge as a result of the time lag between diagnosis and infection. The algorithm estimates the number of infected people at any point of time using the diagnosis time of the individuals, community population, and expected time lag between diagnosis

and infection. The estimation can be done using the equation 3.6.

3.3 Growth-based Ordering with Correction

As discussed earlier, the correction algorithm has been presented to modify the lack of information in the transmission history of the network. This lack of information is due to the time lag between the diagnosis and infection for any infected individuals. Also, the growth-based approach presents a method to sort communities based on their respective growth rate and solve prioritization problem.

One of the challenges for implementing growth-based approach is this lack of information caused by the time lag. Therefore, the growth-based approach can be implemented using the correction algorithm so that the time lag challenge may be addressed to some extent. Therefore, instead of using $D_c(t)$ in the equation 2.1, the estimated the number of infected people $\hat{I}_c(t)$ is applied to equation 2.1 for computing growth rate of communities and their growth-based ordering.

Let $\hat{I}_c(t)$ denote the estimated the number of infected people in community c at time t using the equation 3.8. In this case, the growth rate will be defined as follows:

$$G_c^{\hat{I}}(t-\Delta t, t) = \frac{\hat{I}_c(t) - \hat{I}_c(t-\Delta t)}{\hat{I}_c(t-\Delta t)} \tag{3.8}$$

where $\hat{I}_c(t)$ is the estimation of the number of infected people in the community c at time t. This estimation can be performed using the suggested correction algorithm.

In the following chapters, we will see if the correction over the number of diagnosed people improves the performance of our growth-based approach or not.

3.4 Summary

One of the main challenges for implementing the growth-based approach is the lack of information in the number of infected individuals. This challenge is due to the time lag between diagnosis and infection for any infected individual. This may cause significant inaccuracy for solving prioritization problem and finding the top-growing communities.

We present a simple correction algorithm to ameliorate this lack of information to some extent. The correction algorithm estimates the number of infected people at any point of time using the diagnosis time of the individuals, community population, and expected time lag between diagnosis and infection. Therefore, applying both correction algorithm and growth-based approach to the prioritization problem may improve the accuracy of prioritization. The empirical results of applying the correction algorithm to growth-based approach are discussed in the following chapters.

Chapter 4

Experimental Setup

4.1 FAVITES Simulation Procedure

As discussed in the previous chapters, simulations have been conducted to analyze the growth-based approach and correction algorithm performance for identifying the top growing communities among all of them. Also, the simulations are performed to explore the performance of the aforementioned methods to solve the prioritization problem in different settings of the network.

FAVITES [18] is a modular framework for simulating the epidemic in different settings of the network. The simulation procedure consists of series of interactions between abstract modules. Implementation of modules is user-specified in the sense that users can specify the implementation regarding the context for which they are applying FAVITES.

In this context, FAVITES has been used to model the spread of HIV throughout a con-

tact network of individuals. It consists of steps to construct the contact network, select the initial seeds, and simulate the transmissions between individuals until the user-specified end criteria. FAVITES workflow will be discussed in the following for our specific application. It is worth noting that FAVITES can be applied to a wide range of epidemiological applications although we will talk about the application of FAVITES in our work.

Step 1. The ContactNetworkGenerator module generates a graph representing a contact network. Nodes of the graph represent the individuals and the edges between the nodes represent the possible contact between individuals. Potential transmissions will occur between the individuals which are connected. The graph can be constructed using different stochastic models for the contact network including the Erdős–Rényi model [23], the random partition model [24], the Barabási–Albert model [25], the Caveman model [26], the Watts-Strogatz model [27], and stochastic Block Barabasi-Albert model which will be elaborated on in details in Appendix A . All of the models are implemented in FAVITES.

Step 2. The SeedSelection module initializes the HIV spread procedure with some certain infected nodes. The process is initiated at time zero of the simulation by choosing the individuals who are initially infected. There are different models implemented in FAVITES for seed selection including Random Selection and Edge-Weighted selection.

Step 3. After HIV spread initialization, a series of transmission events will occur until the user-specified end criterion is met. The TransmissionTimeSample module chooses the time of the next transmission. Also, the TransmissionNodeSample chooses a source node and a target node for the next transmission event. A series of iterative transmission events will occur until the user-specified end criteria. The epidemiological model of HIV

transmissions assumes that the individuals will start in a state of a Markov model and transition between different states.

4.2 Simulation Parameters

We have conducted the simulations using a set of base simulation models and parameters. As indicated before, we have explored the performance of proposed methods in different settings of the network. Indeed, we have a set of varying parameters for which the process is simulated. We have ran 20 replicates for each set of parameters.

Contact Network. We have generated a contact network of 200,000 individuals using FAVITES. The contact network consists of 400 communities each with 500 individuals and each of the communities has been generated using stochastic block Barabasi-Albert (BA) model. This is a model we designed and have explained in detail in Appendix A. In this model, each of the communities are modeled as Barabasi-Albert graph with parameters chosen from a predefined distribution. The reason to choose Barabasi-Albert model is its power law degree distribution [25] as sexual networks are assumed to have this property [28]. Also, the design of our specific stochastic block BA model is chosen such that the whole generated contact network maintains the scale-free property of the BA models. The expected degree of the contact network (E_d) has been set to 4 edges. Also, there is another parameter (p_{across}) which is the probability that a given possible inter-community edge is created.

Seeds. The seeds are generated with a geometric distribution. The reason to choose this

distribution is that it is desirable to have no seed in a significant number of communities which is applied to 80 communities out of the whole 400 communities. The average number of seeds in each community equals to 25. The seeds are generated with a geometric distribution with the parameter $p = 0.032$ because we wanted to set 20 percents of communities with zero seed and the average number of seeds per community to 25.

epidemiological model. HIV transmission has been assumed to be a Markov chain model with five different states Susceptible (S), Acute HIV Untreated (AU), Acute HIV Treated with ART (AT), Chronic HIV Untreated (CU), and Chronic HIV Treated with ART [18]. This is a simplified version of the mathematical model proposed by Granich *et al.* [20].

The parameters $\lambda_{AU\to CU}$, $\lambda_{AT\to CT}$ have been set such that the expected transition time from AU to CU is 6 weeks [29] and expected transition time from AT to CT is 12 weeks [30]. The parameter $\lambda_{U\to T}$ has been set such that the expected time between infection and diagnosis (expected time to start ART) equals to 1 year [31]. E_{ART} represents the expected time lag between diagnosis and infection. It is defined as $E_{ART} = \dfrac{1}{\lambda_{U\to T}}$. Finally, the parameter $\lambda_{T\to U}$ is chosen to make the expected transition time between T and U be equal to 25 months [32]. Also, the infection rates $\lambda_{S,AU}$, $\lambda_{S,CU}$, $\lambda_{S,AT}$, $\lambda_{S,CT}$ have been chosen such that $\lambda_{S,AU} = 5\lambda_{S,CU}$ [33], $\lambda_{S,CT} = 0$, and $\lambda_{S,AT} = 0.05\lambda_{S,CU}$ [30]. Finally, $\lambda_{S,CU}$ is chosen to be 0.1 per year.

Varying Parameters. In this work, we mainly focus on two parameters and explore the performance of the proposed methods in different settings of the network corresponding to these two parameters. We varied the expected degree of the contact network in the range 2,4, 8 and examined the networks with different degrees of connectivity. For each degree

of connectivity, we explored E_{ART} from 1/2 to 4. The focus of our work will be on E_{ART} which can be influenced by health departments. Health departments can try to decrease this expected time lag between diagnosis and infection which has significant impact on our understanding of transmission history and contact network.

Table 4.1 represents all the chosen parameters in the simulation procedure.

Parameter	**Parameter Values**
Contact Network Model	**Stochastic Block Barabasi-Albert Model**
Expected Degree (E_d)	2, **4**, 8
Expected Time to ART ($E_{ART} = 1/\lambda_{U \to T}$)	0.5, **1**, 2, 4 (years)
Number of Communities (M)	**400**
Number of Individuals in a Community (N)	**500**
p_{across}	**0.000001**
Average Number of Seeds in a Community	**25**
Number of Communities with No Seed	**80**
$\lambda_{AU \to CU}$	**8.667** (1/years)
$\lambda_{AT \to CT}$	**4.333** (1/years)
$\lambda_{T \to U}$	**0.48** (1/years)
$\lambda_{S,AU}$	**0.5** (1/years)
$\lambda_{S,CU}$	**0.1** (1/years)
$\lambda_{S,AT}$	**0.005** (1/years)
$\lambda_{S,CT}$	**0** (1/years)

Table 4.1. Simulation parameters (base parameters in bold)

4.3 Step-wise Prioritization

As discussed in the previous chapters, we are proposing methods to solve prioritization problem which is defined as finding top growing communities in a network of connected communities. We presented growth-based approach for solving the prioritization problem and addressed the challenges that have been encountered. More specifically, a correction algorithm has been proposed to modify the lack of information caused by the time lag between diagnosis and infection of the individuals.

We presented a real-time method for sorting the communities and simulated the HIV spread in a contact network. The proposed growth-based approach is implemented as well in different settings of the network in order to evaluate the increased ability obtained from the above-mentioned method. Also, we have access to the full information and we can evaluate the performance of the presented growth-based method by determining the optimal solution to the prioritization problem. Performance of the correction algorithm in rectifying the effects of the time lag between diagnosis and infection has also been evaluated. We will introduce evaluation metrics in section 4.4.

As mentioned earlier, the proposed method is a real time method and it can be implemented at any point of time. We have implemented this growth-based approach each six months for the whole network using the partial information that we have from past six months. Thus, we consider the information from past six months to make predictions for the following six months. We start from time zero and implement the growth-based approach and evaluate its statistical performance for solving the prioritization problem every six month. Δt is assumed to be 0.5 year in the equations 2.1 and 2.2.

4.4 Evaluation Metrics

As previously discussed, the goal is to find the top growing communities to capture whether monitoring communities of individuals with the growth-based approach can help health departments to identify the communities with higher risk of transmissions or not.

Assume that the public health intervention efforts will be able to pick a certain number of communities and allocate their limited medical resources to the top-growing communities every six months. Let C_p denote the number of communities which can be picked every six month. After picking top C_p communities out of all communities with the growth-based approach, we will assess the ability of growth-based approach by evaluating its predictive power in the *subsequent* six months. We will mainly consider the best possible ordering of communities as we have the information from future transmissions in our simulation and evaluate the proposed method's performance by comparing its performance to the optimal and random ordering cases.

A metric should be defined to assess the ability of growth-based approach. The metric should represent the predictive power of the proposed method by comparing ordering S as the real-time output of growth-based approach with optimal and random ordering. We will introduce different metrics to evaluate the performance of our proposed methods.

Score. We define the future growth rate for the communities which have been identified as top-growing communities. Assume that C_p communities have been picked as top-growing communities at time t and we want to define a metric to measure the picked communities' growth. We will show the growth rate for the communities as $g(S, C_p, t)$ which refers to

the growth for the whole C_p communities based on the ordering S.

Let's denote $I_p(t)$ as the cumulative number of infected people in all of the C_p communities at time t. We will define the growth rate as follows:

$$g(S, C_p, t) = \frac{I_p(t + \Delta t) - I_p(t)}{I_p(t)} \tag{4.1}$$

This metric gives the average number of new transmissions from each of the infected individuals in the C_p communities that we have picked. This metric can also be expressed as the number of new infections per capita for the targeted communities. It shows how much successful we are in determining the communities with high risk of transmitting. This metric is proposed by Wertheim *et al.* [34] which is based on a practice used on real data obtained from New York City public health HIV-1 surveillance registry.

Kendall Rank Correlation. Kendall rank correlation coefficient is commonly used to measure the ordinal association between different orderings of data [35]. We can say intuitively that the Kendall rank correlation is a measure of similarity between two orderings. If the orderings have high Kendall rank correlation, they will probably have a similar rank. Conversely, if they have low rank correlation, they will be assumed to have dissimilar rank. Assume that we have two orderings $S = (S_1, S_2, S_3, ..., S_M)$ and $R = (R_1, R_2, ..., R_M)$. on a given set of elements $C = \{C_1, C_2, ..., C_M\}$. Consider all of the pairs which can be selected from the set of elements. There are $\binom{M}{2}$ possible pairs to choose from the set of elements. Assume that we have selected (C_i, C_j). Let's denote the indices that C_i and C_j are present in the ordering S, as i_s and j_s. Similarly, let's denote the respective indices in the ordering R as i_r and j_r. We say that the the pair (C_i, C_j) is concordant with respect to the two

orderings S and R if the rank for the two elements agrees. This means that the pair is concordant if both $i_s < j_s$ and $i_r < j_r$; or if both $i_s > j_s$ and $i_r > j_r$. Conversely, the pair is said to be discordant with respect to the two orderings S and R if the rank for the two elements does not agree. This means that we should have either both $i_s < j_s$ and $i_r > j_r$; or both $i_s > j_s$ and $i_r < j_r$.

The Kendall rank correlation coefficient is defined based on the number of concordant and discordant pairs from all of the possible pairs of elements in a given set of elements C with respect to the two orderings $S = (S_1, S_2, S_3, ..., S_M)$ and $R = (R_1, R_2, ..., R_M)$. Therefore, it can be formulated as follows:

$$\tau = \frac{\text{Number of Concordant Pairs} - \text{Number of Discordant Pairs}}{\binom{M}{2}} \tag{4.2}$$

where τ is the Kendall rank correlation coefficient which is applied to the two ordering S and R. It can also be interpreted as the similarity between the two orderings. If the two orderings have similar (ideally identical) rank, we expect τ to be close to 1. However, if the two rankings share a small number of concordant pairs, it seems that the two orderings do not share the same ranking for most of the objects. Therefore, we expect τ to be close to -1. We can view Kendall rank correlation coefficient as a regular correlation coefficient which indicates the similarity between the two orderings. The metric can be used to measure similarity between the output ordering of proposed methods and optimal ordering.

4.5 Summary

Simulations have been conducted to explore the performance of growth-based approach and correction algorithm for solving the prioritization problem. The HIV spread process has been simulated using a modular framework known as FAVITES. The simulation starts at time zero and continues until all of the individuals in the network are infected.

There are two types of parameters in the simulation: fixed and varying parameters. Fixed parameters have been set to specific values. Conversely, there are two varying parameters to explore the performance of the above-mentioned methods in different settings of the network. We will mainly focus on E_{ART} which is defined as the expected time lag between initial infection and diagnosis. This parameter is the main cause of the lack of information about the number of infected people in the communities.

Chapter 5

Results

5.1 Evaluation Details

As discussed in the previous chapters, we have proposed a growth-based approach to identify the top growing communities among all of them in a network in which the HIV is spreading. We have simulated the HIV spread in a network of communities, and implemented the growth-based approach. Also, evaluation metrics have been introduced to evaluate the performance of proposed growth-based approach.

In this approach, a real-time method has been suggested to sort the respective communities. We have evaluated this real-time method based on the real-time ordering produced by the growth-based approach. In order to evaluate the performance of this approach, we should have a benchmark for sorting the communities. Fortunately, we can have optimal ordering at any point of time based on the full information which is obtained from the

transmission events of the network. For evaluation purposes, we will use the information from future transmissions to obtain the optimal ordering at any point of time. We will assess the ability of network-based statistics by comparing our proposed method of ordering communities to the random ordering and optimal ordering of communities.

Figure 5.1 plots the score of different growth-based approaches in different settings of the contact network. As it can be observed, the simulation and evaluation has been conducted for different settings of the contact network. As discussed earlier, we explore the effects of two varying parameters (E_{ART} and E_d) on the proposed growth-based approach. The rows correspond to different connectivity levels of the contact network. The expected degree which can be considered as a measure of network connectivity takes the values 2,4,8 in the simulations. Also, the settings of the network and transmission procedure has changed via the expected diagnosis time (E_{ART}) parameter. E_{ART} is varying from 0.5 year to 4 years which means the expected time lag between diagnosis and infection is changing from 6 months to 4 years. Columns correspond to different values of E_{ART} for which the HIV spread network has been simulated.

Consider one of the subplots of Figure 5.1. We will explain the results of one subplot and then analyze the performance for different settings of the network.

In each of the subplots, the X-axis is C_p which is the number of communities that we pick every time which was introduced in equation 4.1. We are varying the number of communities that we can pick. Also, Y-axis represents the average of the evaluation metric $g(S, C_p, t)$ from $t = 2$ to $t = 10$.

In each subplot, 5 different curves have been plotted: diagnosis, infection, optimal,

correction and random curves.

Random Curve: The random curve corresponds to the random ordering for different values of C_p. Here, instead of proposing methods to sort communities at any point of time, we will sort them randomly and top-growing communities will be selected randomly. This means that the C_p communities will be chosen completely randomly. Then, the score of random ordering will be evaluated. The score of random ordering is expected to be equal the average number of new infections per capita across all of the individuals. Also, the random ordering performance does not change by varying the C_p as the the average number of new infections per capita in the whole network will not be changed for different values of C_p.

Optimal Curve: The optimal curve relates to the optimal ordering for different values of C_p. The optimal ordering refers to the most effective ordering of communities at any point of time. As we have the information on the future transmission events, we will be able to find the best possible ordering in the sense that the ordering will lead to the maximum possible value of the evaluation metrics at any point of time. Therefore, the ground truth, which is the best possible ordering, is accessible. The score of the optimal ordering has been plotted by varying the C_p values.

Infection Curve: The infection curve refers to the ordering resulted from growth-based approach by using full transmission history. For implementing growth-based ordering to obtain the infection curve, it is assumed that we have the full history about the diagnosis and infection of the individuals in the network to assess the ideal performance of growth-based approach. As mentioned earlier, we don't have access to the full transmission history

since there is always a time lag between diagnosis and infection of the individuals in the HIV spread network. We have implemented the infection growth-based approach to evaluate the performance of our proposed method to address the first challenge which was explained in the section 2.2. Indeed, we wish to evaluate the predictive power of our proposed growth-based approach disregarding the time lag challenge explained in section 2.2. Specifically, this approach has been implemented based on the equation 2.2 although it cannot be implemented in reality.

Diagnosis Curve: The diagnosis curve corresponds to the ordering using growth-based approach based on diagnosis dasta. It is assumed that we only have access to the diagnosis times of infected individuals not their infection times. Therefore, we are evaluating the performance of growth-based approach when the diagnosis times are used for real-time ordering of communities. Indeed, diagnosis curve refers to the implementation of equation 2.1 for the growth-based approach. It is worth noting that correction algorithm has not been implemented for this curve.

Correction Algorithm Curve: Finally, the correction algorithm represents the ordering resulted from applying correction algorithm to the growth-based approach. As it was discussed, we came up with the correction algorithm method to modify the lack of information caused by the time lag between diagnosis and infection. To be more clear, this curve relates to the implementation of equation 3.8 for various values of C_p. The performance of this implementation indicates the growth-based approach performance to address both of the challenges discussed in section 2.2.

Figure 5.1 gives us a curve where higher is better. Furthermore, we may need a single

value for comparing different methods. This means that instead of a curve for different values of C_p, we wish to represent the performance of different methods with just one specific value. As it can be observed in Figure 5.1, there are three values in each of the subplots: Green, red, and blue values which relates to the green, red, and blue curves respectively. We will also explain the formulation for computing these values. Let a_M denote the area between curve of method M and random method's curve in Figure 5.1. The values are computed using the following formula:

$$\text{effectiveness of diagnosis approach} = \frac{a_{Diagnosis}}{a_{Optimal}}$$

$$\text{effectiveness of infection approach} = \frac{a_{infection}}{a_{Optimal}}$$

$$\text{effectiveness of correction algorithm approach} = \frac{a_{correction\ algorithm}}{a_{Optimal}}$$

We will refer to these values as the effectiveness of implemented methods which indicates the ability of the methods to identify the top growing communities. Indeed, this value is a fraction between 0 and 1. If the value is 1, then the performance of the implemented method is the best possible performance. If the value equals to 0, the method is performing no better than a random ordering and does not contain any ability.

5.2 Evaluating Growth-based Approach

Before evaluating the growth-based approach, we will explore some general trends which can be observed in each of the subplots of the Figure 5.1. Considering all of the subplots

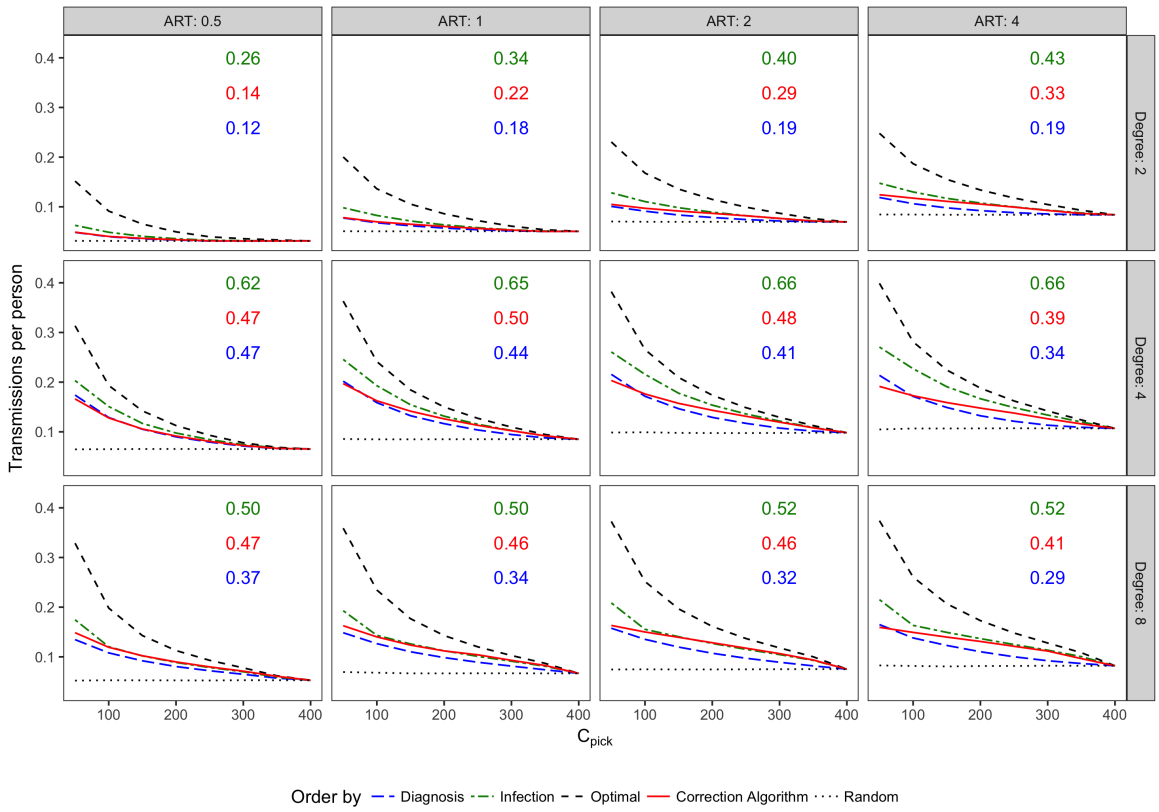

Figure 5.1. The score for different methods of sorting has been evaluated. The score is defined as the number of new infections within the time slot $\Delta t = 0.5$ year from the current time t. The score has been defined in equation 4.1 which is referred as the number of new infections per capita. The performance of proposed growth-based approach has been evaluated by comparing its score with the score of random and optimal ordering. The performance of implemented correction algorithm has been evaluated in modifying the time lag between infection and diagnosis.

in the Figure 5.1, the score strictly decreases as C_p increases. This trends makes sense since by increasing C_p, we will pick more communities at any point of time. Therefore, the average infectiousness of the selected communities will be decreased. As a result, we expect to have less number of new infections per capita as C_p increases. Also, when

$C_p = 400$, the performance of all of them will converge to the same value (average number of new infections per capita in the whole network) which is obviously expected as all of the communities are selected and there will be no difference between different approaches.

It can also be observed generally in all of the subplots that there is always a difference between performance of different methods. The infection time approach is closest to the optimal ordering, but is not possible to be implemented in reality as infection times are not known. Then, we have the correction algorithm approach which is performing somewhere between infection time approach and diagnosis time approach. The diagnosis time approach will come next. Finally, the random ordering comes which does not include any ability to prioritize.

Now, we aim to analyze the performance of growth-based approach disregarding the time lag challenge. The time lag challenge and performance of correction algorithm to address the time lag challenge will be discussed in the section 5.3.

If we take a closer look at the green values in the Figure 5.1 which corresponds to growth-based approach using the full transmission history (infection curve), we can observe its performance is somewhere between optimal ordering and random ordering. It performs up to 60% of the performance of best possible ordering. In some cases, its performance goes down to 26% of the optimal ordering. These results indicate that the transmission history may have valuable information about the future transmission events.

Also, if we compare subplots in different columns, it can be inferred that the growth-based approach for the full transmission history will perform closer to the optimal ordering as E_{ART} increases. Clearly, if we increase the expected time lag between diagnosis and

infection, we expect the epidemic to spread more rapidly as we will have more individuals who are infected but not under treatment (individuals in AU state will have the highest rate of infectiousness). Therefore, the growth-based approach using the full transmission history will perform slightly better for the large values of E_{ART}. This pattern makes sense due to the fact that by increasing values of E_{ART}, there will be more distinction between different communities. This distinction arises as E_{ART} increases because HIV will spread more rapidly in the most infectious communities and they will be identified as the top-growing communities comparing to the uninfected or the ones which have less number of initial infections.

If the subplots in different rows are compared, the performance of growth-based approach versus connectivity levels in the network goes up first and then decreases as the connectivity level increases. The best performance for different degrees of connectivity refers to the connectivity level of $E_d = 4$. It might be because of the fact that increasing connectivity levels will make the HIV spread quicker. Therefore, we expect to have better performance since there will be more distinction between communities as described earlier. On the other hand, if the connectivity level is too high, it is expected that the number of infections in highly infectious communities will be saturated rapidly and there will be less space for improvement in predicting the future transmission. Indeed, all the individuals of highly infected communities will be infected in a relatively small period of time which makes our proposed growth-based approach performs worse comparing to the lower connectivity levels of the network.

5.3 Evaluating the Correction Algorithm

In order to explore the effects of the time lag on the suggested approach, we take a closer look at the blue curves. The values in the Figure 5.1 indicate the performance of the growth-based approach using diagnosis times of individuals as implemented using the equation 2.1. This comparison will shed light on the encountered time lag challenge.

Considering the diagnosis curves in the subplots of the Figure 5.1, the diagnosis approach is performing worse than the infection time in different settings of the network. In most cases, diagnosis approach performs with a performance around half of the performance of the infection approach. In some cases, the performance is less than half of the infection approach performance. This observation confirms the fact that lack of information impacts the performance of our growth-based approach to identify the top-growing communities.

Also, if we compare the subplots in different columns for the diagnosis growth-based approach, a general trend is observed. The gap between the performance of growth-based approach (implemented using the diagnosis time of the individuals) is more clear when E_{ART} increases. This pattern is expected since increasing E_{ART} increases the time lag between diagnosis and infection of the individuals. Indeed, the more the E_{ART} is, the more lack of information we have in our growth-based approach for sorting communities. For instance, consider the subplots in the Figure 5.1 for $E_d = 4$. As E_{ART} increases, the diagnosis growth-based approach is performing worse although the performance of infection approach is better for larger values of E_{ART}. Therefore, we expect to have a considerable gap between performance of diagnosis growth-based approach and infection growth-based

approach as the expected time lag between diagnosis and infection increases.

If the subplots for different rows of the Figure 5.1 are compared, it can be inferred that the difference between performance of infection and diagnosis approach is reduced for networks with high level of connectivity. As discussed earlier, if the connectivity level is relatively high, it is expected that the number of infections in highly infectious communities will be saturated rapidly and there will be less space for improvement in predicting the future transmission events. Therefore, the effect of the time lag on the performance of different approaches for highly connected networks won't be as significant as the difference in the networks with low level of connectivity.

Now that we have explored the effects of the time lag on the performance of our growth-based approach, we evaluate the performance of correction algorithm which is proposed to address the time lag challenge between diagnosis and infection of the individuals.

To evaluate the performance of correction algorithm, we will delve into the performance of correction algorithm in different settings of the network in the Figure 5.1. As it can be observed in each of the subplots of this figure, implementing the correction algorithm will address the time lag challenge to some extent. In some certain settings, a significant improvement is observed for addressing the time lag challenge. For instance, consider the subplot for $E_d = 8$ and $E_{ART} = 1$. The effect of the time lag challenge for implementing the growth-based approach may be considered as the difference between the green value and blue value in the figure corresponding to infection curve and diagnosis curve, respectively. This difference between the two values which is equal to 0.16 represents the difference in performance of diagnosis and infection approach for growth-based ordering. However, the

performance of correction algorithm (red value) is 0.46. This observation confirms that implementing correction algorithm will improve the growth-based ordering in some cases up to 75% ($\frac{0.46 - 0.34}{0.5 - 0.34} = 0.75$).

Comparing the subplots for different columns in the Figure 5.1 and considering the performance of correction algorithm, we may conclude that the effect of implementing correction algorithm will be more noticeable as the expected time lag between diagnosis and infection increases. This observation points to the fact that the correction algorithm will have more space for improvement in the networks with relatively large time lag between diagnosis and infection. Indeed, as the time lag between diagnosis and infection increases, we expect our proposed correction algorithm to perform better since the time lag challenge will cause significant amount of lack of information.

Also, if we compare the performance of the correction algorithm in addressing the time lag challenge for different connectivity levels of the network in the Figure 5.1, we observe that the improvement of correction algorithm on highly connected networks is more substantial. As the connectivity level increases, it is expected that the number of infections in highly infectious communities will be saturated rapidly. This means that the first challenge which is considered as the difficulty of using transmission history for future transmission prediction may not be addressed efficiently. In this case, the second challenge which is attempted to be addressed by correction algorithm plays an important role. Therefore, there will be more space for addressing this challenge in the highly connected networks.

5.4 Sensitivity Analysis

In the previous sections, we have evaluated the performance of growth-based approach and correction algorithm to address the challenges mentioned in section 2.2. We observed that correction algorithm improves the performance of growth-based ordering in different settings of the network.

However, there are two assumptions for implementing the correction algorithm that can be challenged. In order to implement the correction algorithm, we used equation 3.7. In this equation, it is assumed that E_{ART} is known although it should be estimated. In order to show that this assumption does not impact the performance of correction algorithm, a sensitivity analysis has been performed on E_{ART}. The results of the sensitivity analysis on E_{ART} has been shown in the Figure 5.2.

In each of the subplots, the correction algorithm has been evaluated for the cases that the estimated value for E_{ART} is half or double of its real value. The blue curve (value) in Figure 5.2 corresponds to the implementation of correction algorithm with the estimation equal to half of the real value of E_{ART}. The green curve (value) refers to the implementation with double value estimation of E_{ART}.

As it is observed in the Figure 5.2, the performance of the correction algorithm remains relatively robust to the inaccurate estimations of E_{ART}. The change in the performance compared to the optimal settings remains unchanged in some settings (e.g., ART:4, Degree: 8), and changes slightly in others (e.g., ART:1, Degree:8). The most reduction in accuracy is for Degree 8, ART: 0.5, where inaccurate estimates of E_{ART} can reduce the performance by 6% (e.g., from 47% to 41%).

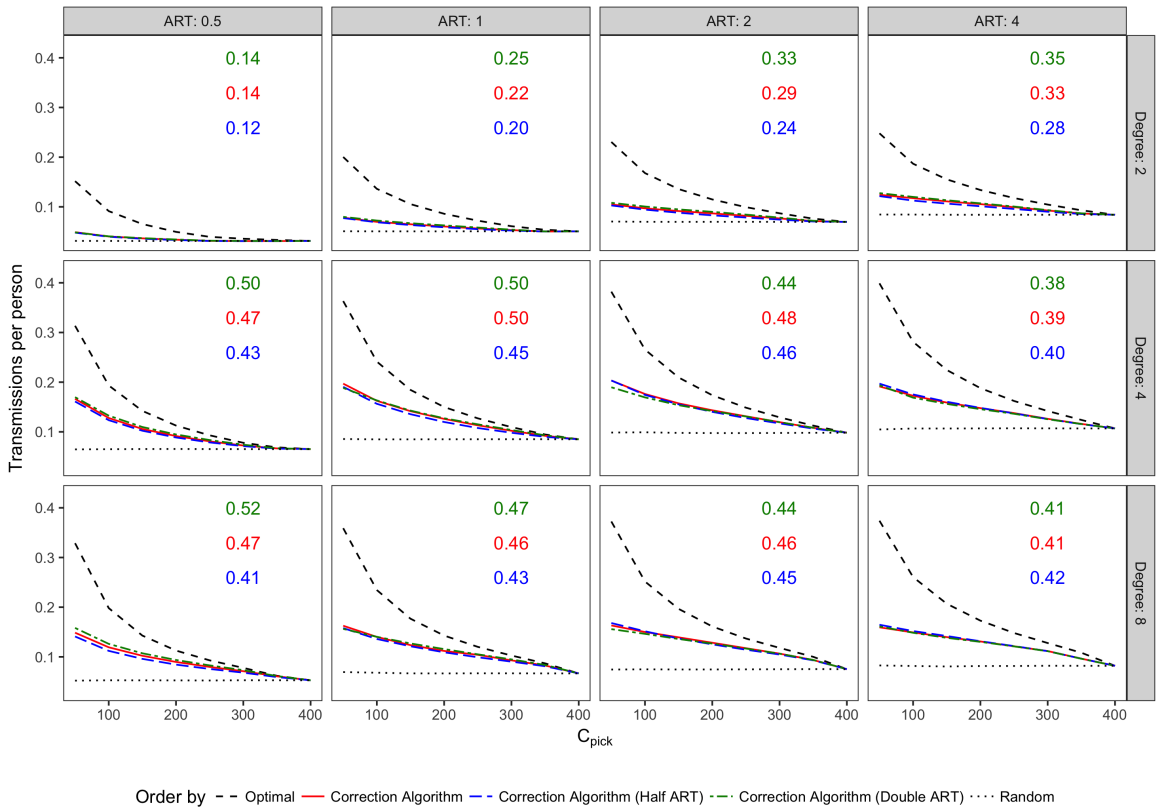

Figure 5.2. Sensitivity analysis on the correction algorithm score for half or double estimations of E_{ART}. The score has been defined in equation 4.1 for $\Delta t = 0.5$ year which is referred as the number of new infections per capita. The performance of correction algorithm has been evaluated with wrong estimations of E_{ART}.

Furthermore, the number of individuals in each of the communities ($N = 500$) is also assumed to be known accurately for implementing correction algorithm. A sensitivity analysis has also been done for the cases that the estimation of communities' population is half or double of the correct value of communities' population. The results of sensitivity analysis for the change in communities' population has been shown in the Figure 5.3. Once again, the correction algorithm seems robust to wrong estimations for this parameter. The

maximum reduction in accuracy is 3%, which again happens for Degree: 8 and ART: 0.5.

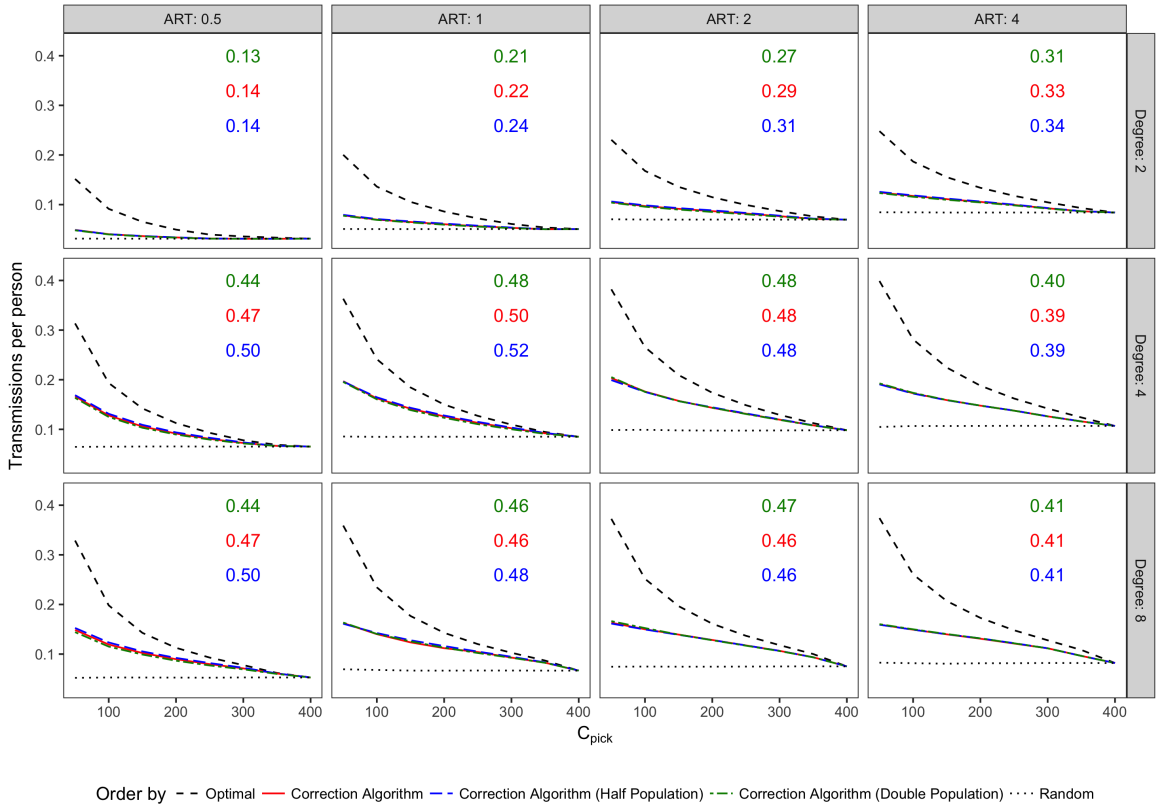

Figure 5.3. Sensitivity analysis on the correction algorithm score for half or double estimations of N which is each of the communities' population. The score has been defined in equation 4.1 for $\Delta t = 0.5$ year which is referred as the number of new infections per capita. The performance of correction algorithm has been evaluated with wrong estimations of N.

In conclusion, results of the correction algorithm sensitivity analysis confirms the robustness of correction algorithm to the wrong estimations of parameters (E_d and N) in various settings of the network.

5.5 Comparison on Different Approaches

We have evaluated the performance of growth-based approach and correction algorithms in the sections 5.2 and 5.3. The next step is to compare different methods with performance metrics which will be defined carefully.

Consider the the correction algorithm, infection, diagnosis, random, and optimal curves in the Figure 5.1. We will define the performance metric to evaluate each of the methods for different values of C_p for infection, diagnosis, and correction algorithms curve. Let's denote the correction algorithm, infection, diagnosis, random, and optimal curves as $E(C_p)$, $I(C_p)$, $D(C_p)$, $R(C_p)$, and $O(C_p)$ respectively. The performance of each of diagnosis, infection, and correction algorithms will be defined as:

$$\text{effectiveness of diagnosis curve} = \frac{D(C_p) - R(C_p)}{O(C_p) - R(C_p)}$$

$$\text{effectiveness of infection curve} = \frac{I(C_p) - R(C_p)}{O(C_p) - R(C_p)}$$

$$\text{effectiveness of diagnosis curve} = \frac{E(C_p) - R(C_p)}{O(C_p) - R(C_p)}$$

The metrics represent the ability of each of the infection, diagnosis, and correction algorithm methods for different values of C_p. Then, this metric has been plotted for different values of C_p in various settings of contact network in the Figure 5.4.

Considering each of the subplots of the Figure 5.4, it can be concluded that the effectiveness of correction algorithm is somewhere between the effectiveness of diagnosis and infection growth-based approaches. Indeed, the respective values for different values of C_p

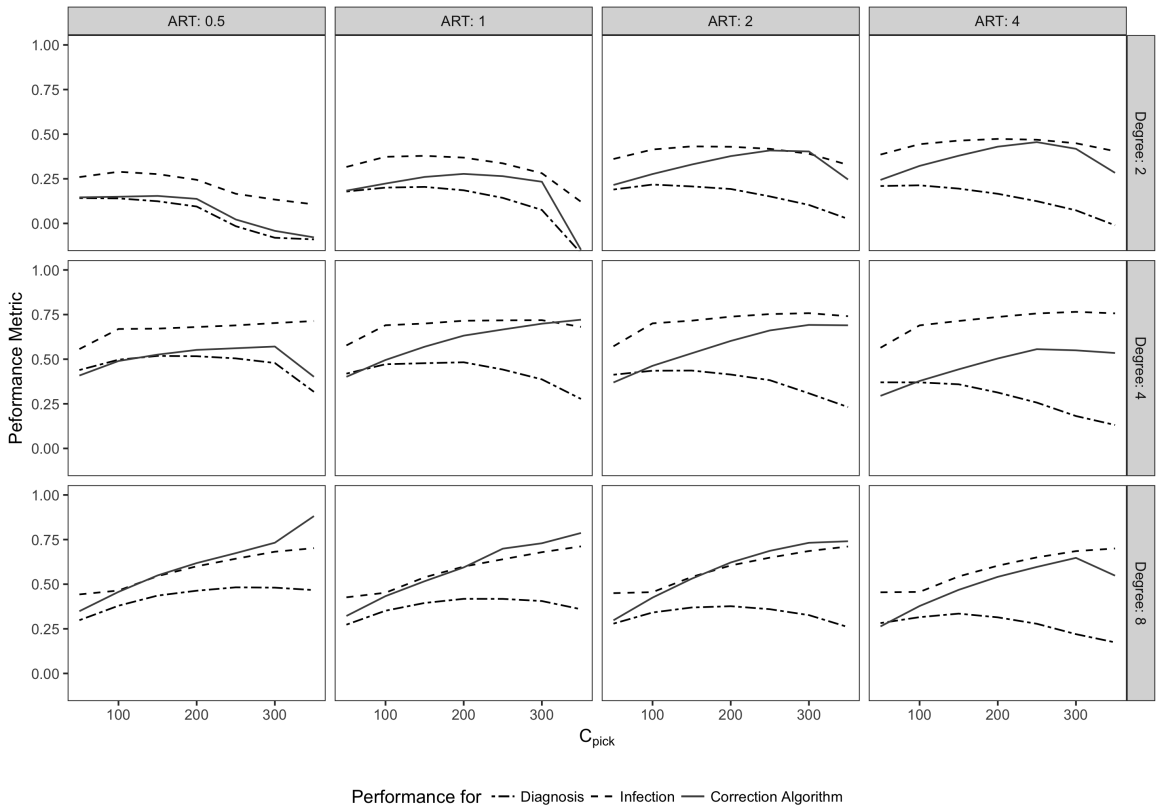

Figure 5.4. The effectiveness of each of the infection, diagnosis, and correction algorithm approaches. The effectiveness has been computed for different values of C_p. The performance of proposed growth-based approach for diagnosis, infection, and correction algorithm has been evaluated by measuring the effectiveness.

indicates that the proposed correction algorithm has been successful in modifying the gap between infection and diagnosis growth-based approaches. Therefore, correction algorithm seems to be an effective method for modifying the lack of information caused by the time lag between infection and diagnosis.

We will also introduce another method for comparing the different growth-based approaches. As discussed earlier, each of the growth-based approaches presents a method for

real-time ordering of communities. In our work, the ordering has been implemented for all of the approaches every $\Delta t = 0.5$ year. Then, based on the output ordering of each of the methods, the Kendall rank correlation coefficient which is introduced in section 4.4 has been computed between the optimal ordering and each of the method's respective ordering. The results for different values of E_{ART} has been presented in the Figure 5.5.

It is also worth noting that Kendall rank correlation coefficient measures the similarity of two orderings. Therefore, the more correlation coefficient is, the more similar the output ordering is to the optimal ordering of communities. Also, from the definition of this metric in the equation 4.2, it can be understood that the metric does not rely on C_p since in the definition of Kendall rank correlation coefficient, all of the communities are considered not C_p of them as C_p communities of the whole communities are considered for the score metric in the equation 4.1.

If we take a closer look at the subplots of Figure 5.5, it is inferred that the ordering of communities resulted from correction algorithm approach is more similar to the optimal ordering comparing to the output ordering of diagnosis growth-based approach. However, it will not obviously perform as well as infection growth-based approach since in the correction algorithm, the full transmission history is not accessible although it is available for implementing infection-based approach.

In order to evaluate the performance of different methods for different values of E_{ART}, consider each of the columns of Figure 5.5 separately. As E_{ART} increases, we will have a better performance for correction algorithm and infection approach. However, the diagnosis growth-based approach will perform worse as E_{ART} increases. Indeed, by increasing E_{ART},

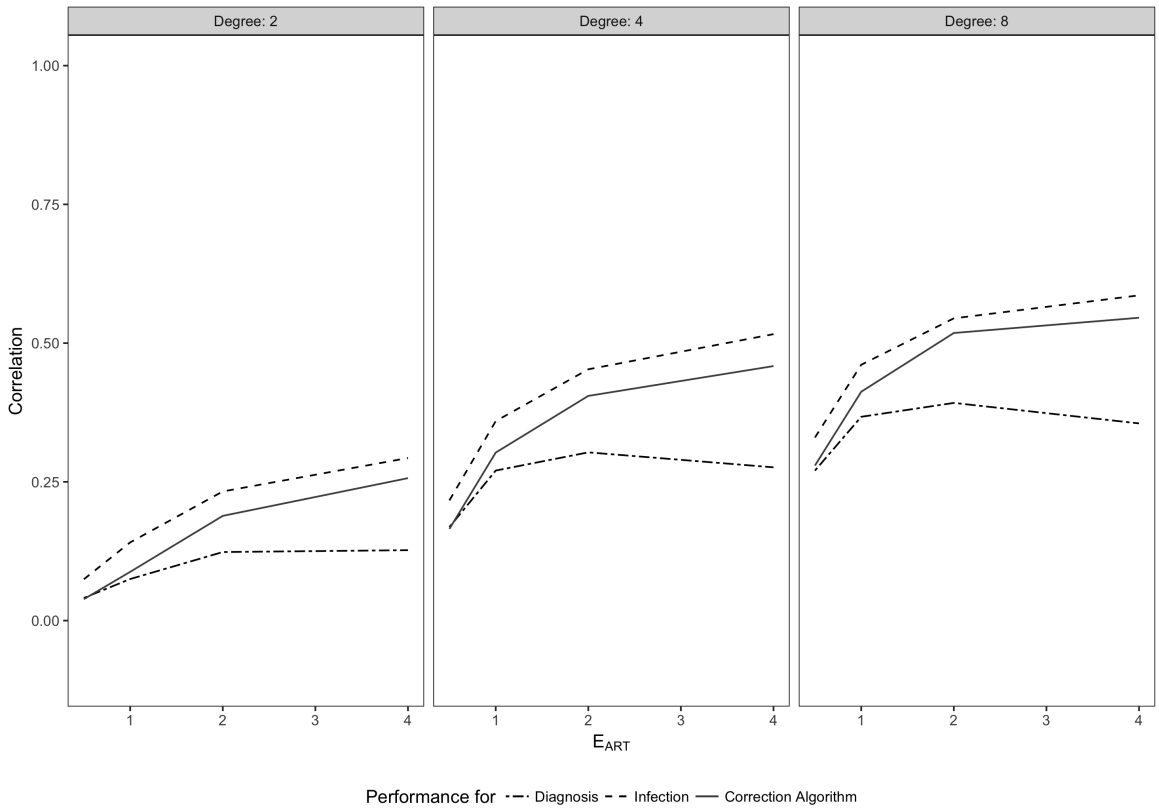

Figure 5.5. The Kendall rank correlation coefficient between the output ordering of different methods and optimal ordering has been computed. The Kendall rank correlation coefficient has been defined in equation 4.2. The performance of proposed growth-based approach has been evaluated by measuring the similarity between its output ordering and optimal ordering which can be computed by Kendall rank correlation coefficient. The performance of implemented correction algorithm has also been evaluated in modifying the time lag between infection and diagnosis.

the time lag between diagnosis and infection increases. Then, there will be more lack of information caused by this time lag. Thus, the diagnosis growth-based approach which is implemented based on diagnosis times of the individuals will not perform well. Fortunately, the performance of our proposed correction algorithm improves which means that the

proposed approach can successfully modify the time lag between diagnosis and infection.

Considering the columns of Figure 5.5, we observe that the correction algorithm can modify the lack of information in the network with various levels of connectivity. If the network is highly connected, there will be more distinction between communities. It might be because of the fact that increasing connectivity levels will make the HIV spread quicker. Therefore, the initially infected communities will be infected quickly, and there will be more distinction between infected and uninfected communities. Then, the performance of the proposed approaches will be better for highly connected networks as depicted in the Figure 5.5.

5.6 Summary

In this chapter, we introduced metrics to analyze the performance of the proposed methods in the previous chapters. Then, the performance of growth-based approach has been evaluated. We have conducted the evaluation for different values of E_{ART} and various connectivity level of the network.

Also, the performance of proposed correction algorithm has been evaluated. Due to the results of evaluation, it is concluded that the proposed correction algorithm is successful in modifying the time lag between diagnosis and infection. Therefore, the lack of information may be rectified to some extent.

Then, an overall comparison is performed between all of the methods which has been implemented. Due to the results in this chapter, it can be concluded that the growth-

based approach together with correction algorithm will have the ability to monitor the HIV spread in the network.

References

[1] J. A. Kelly, J. S. St Lawrence, Y. E. Diaz, L. Y. Stevenson, A. C. Hauth, T. L. Brasfield, S. C. Kalichman, J. E. Smith, and M. E. Andrew, "Hiv risk behavior reduction following intervention with key opinion leaders of population: an experimental analysis.," *American journal of public health*, vol. 81, no. 2, pp. 168–171, 1991.

[2] N. C. H. P. T. Group, "Methodological overview of a five-country community-level hiv/sexually transmitted disease prevention trial.," *AIDS (London, England)*, vol. 21, p. S3, 2007.

[3] J. O. Wertheim, S. L. K. Pond, S. J. Little, and V. De Gruttola, "Using hiv transmission networks to investigate community effects in hiv prevention trials," *PloS one*, vol. 6, no. 11, 2011.

[4] F. Lewis, G. J. Hughes, A. Rambaut, A. Pozniak, and A. J. L. Brown, "Episodic sexual transmission of hiv revealed by molecular phylodynamics," *PLoS medicine*, vol. 5, no. 3, 2008.

[5] B. Vrancken, A. Rambaut, M. A. Suchard, A. Drummond, G. Baele, I. Derdelinckx, E. Van Wijngaerden, A.-M. Vandamme, K. Van Laethem, and P. Lemey, "The genealogical population dynamics of hiv-1 in a large transmission chain: bridging within and among host evolutionary rates," *PLoS computational biology*, vol. 10, no. 4, 2014.

[6] T. Leitner, D. Escanilla, C. Franzen, M. Uhlen, and J. Albert, "Accurate reconstruction of a known hiv-1 transmission history by phylogenetic tree analysis," *Proceedings of the National Academy of Sciences*, vol. 93, no. 20, pp. 10864–10869, 1996.

[7] D. M. Smith, S. J. May, S. Tweeten, L. Drumright, M. E. Pacold, S. L. Kosakovsky Pond, R. L. Pesano, Y. S. Lie, D. D. Richman, S. D. W. Frost, C. H. Woelk, and S. J. Little, "A public health model for the molecular surveillance of hiv

transmission in san diego, california," *AIDS (London, England)*, vol. 23, no. 2, p. 225, 2009.

[8] B. Brenner, M. A. Wainberg, and M. Roger, "Phylogenetic inferences on hiv-1 transmission: implications for the design of prevention and treatment interventions," *AIDS (London, England)*, vol. 27, no. 7, p. 1045, 2013.

[9] S. J. Little, S. L. K. Pond, C. M. Anderson, J. A. Young, J. O. Wertheim, S. R. Mehta, S. May, and D. M. Smith, "Using hiv networks to inform real time prevention interventions," *PloS one*, vol. 9, no. 6, 2014.

[10] J. O. Wertheim, K. Scheffler, J. Y. Choi, D. M. Smith, and S. L. Kosakovsky Pond, "Phylogenetic relatedness of hiv-1 donor and recipient populations," *The Journal of infectious diseases*, vol. 207, no. 7, pp. 1181–1182, 2013.

[11] M. K. Grabowski and A. D. Redd, "Molecular tools for studying hiv transmission in sexual networks," *Current Opinion in HIV and AIDS*, vol. 9, no. 2, p. 126, 2014.

[12] V. Novitsky, S. Moyo, Q. Lei, V. DeGruttola, and M. Essex, "Impact of sampling density on the extent of hiv clustering," *AIDS research and human retroviruses*, vol. 30, no. 12, pp. 1226–1235, 2014.

[13] M. C. F. Prosperi, M. Ciccozzi, I. Fanti, F. Saladini, M. Pecorari, V. Borghi, S. Di Giambenedetto, B. Bruzzone, A. Capetti, A. Vivarelli, S. Rusconi, M. C. Re, M. R. Gismondo, L. Sighinolfi, R. R. Gray, M. Salemi, M. Zazzi, A. De Luca, and on behalf of the ARCA collaborative group, "A novel methodology for large-scale phylogeny partition," *Nature communications*, vol. 2, no. 1, pp. 1–10, 2011.

[14] M. Ragonnet-Cronin, E. Hodcroft, S. Hué, E. Fearnhill, V. Delpech, A. J. L. Brown, and S. Lycett, "Automated analysis of phylogenetic clusters," *BMC bioinformatics*, vol. 14, no. 1, p. 317, 2013.

[15] M. Balaban, N. Moshiri, U. Mai, X. Jia, and S. Mirarab, "Treecluster: Clustering biological sequences using phylogenetic trees," *PloS one*, vol. 14, no. 8, 2019.

[16] S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, and J. O. Wertheim, "Hivtrace (transmission cluster engine): a tool for large scale molecular epidemiology of hiv-1 and other rapidly evolving pathogens," *Molecular biology and evolution*, vol. 35, no. 7, pp. 1812–1819, 2018.

[17] J. O. Wertheim, B. Murrell, S. R. Mehta, L. A. Forgione, S. L. Kosakovsky Pond, D. M. Smith, and L. V. Torian, "Growth of hiv-1 molecular transmission clusters in new york city," *The Journal of infectious diseases*, vol. 218, no. 12, pp. 1943–1953, 2018.

[18] N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, and S. Mirarab, "Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences," *Bioinformatics*, vol. 35, no. 11, pp. 1852–1861, 2019.

[19] A. Cori, H. Ayles, N. Beyers, A. Schaap, S. Floyd, K. Sabapathy, J. W. Eaton, K. Hauck, P. Smith, S. Griffith, A. Moore, D. Donnell, S. H. Vermund, S. Fidler, R. Hayes, C. Fraser, and H. . P. S. Team, "Hptn 071 (popart): a cluster-randomized trial of the population impact of an hiv combination prevention intervention including universal testing and treatment: mathematical model," *PloS one*, vol. 9, no. 1, 2014.

[20] R. M. Granich, C. F. Gilks, C. Dye, K. M. De Cock, and B. G. Williams, "Universal voluntary hiv testing with immediate antiretroviral therapy as a strategy for elimination of hiv transmission: a mathematical model," *The Lancet*, vol. 373, no. 9657, pp. 48–57, 2009.

[21] O. Ratmann, E. B. Hodcroft, M. Pickles, A. Cori, M. Hall, S. Lycett, C. Colijn, B. Dearlove, X. Didelot, S. Frost, A. M. M. Hossain, J. B. Joy, M. Kendall, D. Kühnert, G. E. Leventhal, R. Liang, G. Plazzotta, A. F. Poon, D. A. Rasmussen, T. Stadler, E. Volz, C. Weis, A. J. Leigh Brown, C. Fraser, and on behalf of the PANGEA-HIV Consortium, "Phylogenetic tools for generalized hiv-1 epidemics: findings from the pangea-hiv methods comparison," *Molecular biology and evolution*, vol. 34, no. 1, pp. 185–203, 2017.

[22] J. K. Weltman, "Analytic approximations of sir compartmental models of infectious disease epidemics," *Journal of Medical Microbiology Diagnosis*, vol. 01, no. 04, 2012.

[23] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[24] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[25] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[26] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology*, vol. 105, no. 2, pp. 493–527, 1999.

[27] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[28] D. T. Hamilton, M. S. Handcock, and M. Morris, "Degree distributions in sexual networks: a framework for evaluating evidence," *Sexually transmitted diseases*, vol. 35, no. 1, p. 30, 2008.

[29] S. E. Bellan, J. Dushoff, A. P. Galvani, and L. A. Meyers, "Reassessment of hiv-1 acute phase infectivity: accounting for heterogeneity and study design with simulated cohorts," *PLoS medicine*, vol. 12, no. 3, 2015.

[30] M. S. Cohen, Y. Q. Chen, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, B. Grinsztejn, J. H. Pilotto, S. V. Godbole, S. Mehendale, S. Chariyalertsak, B. R. Santos, K. H. Mayer, I. F. Hoffman, S. H. Eshleman, E. Piwowar-Manning, L. Wang, J. Makhema, L. A. Mills, G. de Bruyn, I. Sanne, J. Eron, J. Gallant, D. Havlir, S. Swindells, H. Ribaudo, V. Elharrar, D. Burns, T. E. Taha, K. Nielsen-Saines, D. Celentano, M. Essex, and T. R. Fleming, "Prevention of hiv-1 infection with early antiretroviral therapy," *New England journal of medicine*, vol. 365, no. 6, pp. 493–505, 2011.

[31] M. O'Brien and M. Markowitz, "Should we treat acute hiv infection?," *Current HIV/AIDS Reports*, vol. 9, no. 2, pp. 101–110, 2012.

[32] B. Nosyk, L. Lourenço, J. E. Min, D. Shopin, V. D. Lima, and J. S. Montaner, "Characterizing retention in haart as a recurrent event process: insights into 'cascade churn'," *AIDS (London, England)*, vol. 29, no. 13, p. 1681, 2015.

[33] M. J. Wawer, R. H. Gray, N. K. Sewankambo, D. Serwadda, X. Li, O. Laeyendecker, N. Kiwanuka, G. Kigozi, M. Kiddugavu, T. Lutalo, F. Nalugoda, F. Wabwire-Mangen, M. P. Meehan, and T. C. Quinn, "Rates of hiv-1 transmission per coital act, by stage of hiv-1 infection, in rakai, uganda," *The Journal of infectious diseases*, vol. 191, no. 9, pp. 1403–1409, 2005.

[34] J. Wertheim, B. Murrell, and L. Forgione, "Torian (2017). cluster growth dynamics suggest strategy for targeted intervention in new york city public health hiv-1 surveillance registry," *HIV Dynamics & Evolution*, no. 1122, p. 38.

[35] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[36] A. Schneeberger, C. H. Mercer, S. A. Gregson, N. M. Ferguson, C. A. Nyamukapa, R. M. Anderson, A. M. Johnson, and G. P. Garnett, "Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in britain and zimbabwe," *Sexually transmitted diseases*, vol. 31, no. 6, pp. 380–387, 2004.

[37] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific american*, vol. 288, no. 5, pp. 60–69, 2003.

APPENDICES

Appendix A

Stochastic Block BA Model

Previous studies suggest that the distribution of number of sexual partners are well-described by power laws and a scale-free network is a suitable model for the sexual contact network [36]. The Barabasi-Albert model would be an appropriate model for generating the sexual contact network [37]. However, in our work, if all of the communities are chosen to be generated by Barabasi-Albert model, the whole contact network may not be scale-free. In order to make the contact network scale free, the communities have been chosen using a predefined distribution.

Our model consists of C number of Barabasi-Albert communities with different parameters of attachment (m). In order to have connection among these networks, we add some random edges between communities with low probability. Then, we want to see the degree distribution for all of the nodes in the whole contact network and see how close it is to a low power distribution.

Suppose that we have C number of BA communities of M different parameters of attachment. Then, we should have a distribution for parameters of attachment of these C communities. In our model, we assume that the probability mass function for parameters of attachment of these communities is as follows:

$$P(m = i) = \frac{2\alpha}{i(i+1)} \tag{A.1}$$

Then, we know that the sum of probability mass function for m from 1 to M must add up to one:

$$\sum_{i=1}^{M} P(m = i) = \sum_{i=1}^{M} \frac{2\alpha}{i(i+1)} = \frac{2\alpha M}{M+1} = 1 \tag{A.2}$$

Therefore, we will have:

$$\alpha = \frac{M+1}{2M} \tag{A.3}$$

Also, we want to calculate the degree distribution (d) for all of the nodes in our whole network. First, we know that the degree distribution in a single BA community with parameter m is as below (when the number of individuals in the community goes to infinity or is very large):

$$P(d = k) = \begin{cases} \frac{2m(m+1)}{k(k+1)(k+2)} & k \geq m \\ 0 & Otherwise \end{cases} \tag{A.4}$$

Now, we can compute the degree distribution for the whole network given the probability mass function of the parameters of attachment (m) and the degree distribution in each community. After doing calculations and using law of total probability, we will have the

following equation for the degree distribution in the whole network:

$$P(d=k) = \begin{cases} \dfrac{4\alpha M}{k(k+1)(k+2)} & k > M \\ \\ \dfrac{4\alpha}{(k+1)(k+2)} & 1 \leq k \leq M \end{cases} \tag{A.5}$$

Also, expected degree of the whole network can be easily calculated as below:

$$E(d) = \sum_{k=1}^{\infty} kP(d=k) = \sum_{k=1}^{M} kP(d=k) + \sum_{k=M+1}^{\infty} kP(d=k) = 4\alpha(-1 + \sum_{k=1}^{M+1} \frac{1}{k}) \tag{A.6}$$

Due to the equation A.3, we know that $\alpha = \dfrac{M+1}{2M}$. Thus,

$$E(d) = 2(\frac{M+1}{M})(H_{M+1} - 1) \tag{A.7}$$

Where H_{M+1} is the sum of the first $M+1$ terms of the harmonic series.

In the figure A.1, the degree distribution of the proposed stochastic block BA model is plotted for different values of expected degree (2,4,8). As it is observed in the figure, the distribution is similar to the scale-free networks proposed for sexual contact networks.

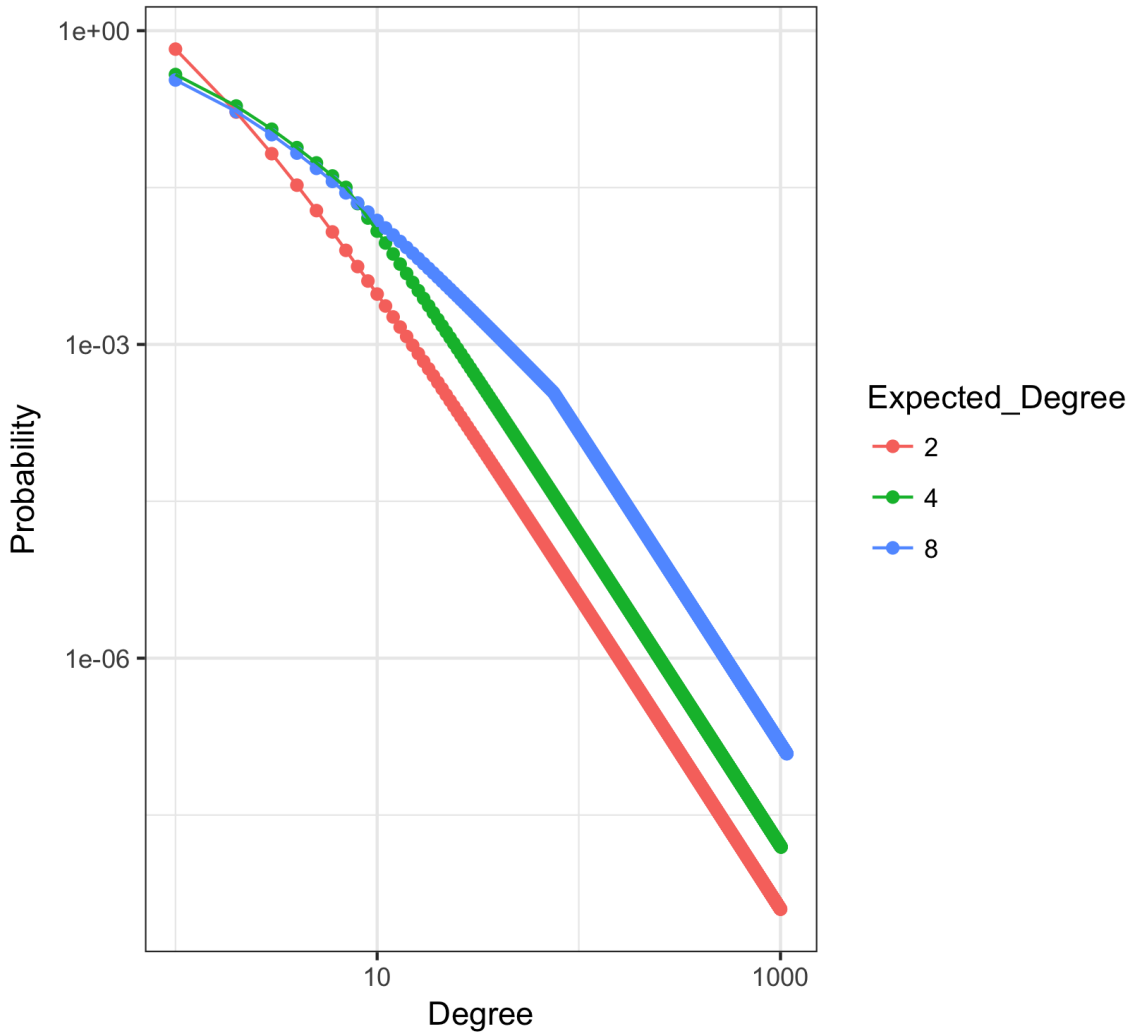

Figure A.1. Degree distribution of proposed stochastic block BA model has been depicted on the log-log scale. As expected, the degree distribution behaves similar to the degree distribution of scale-free sexual networks which follows power law distribution.